Report on

Visualization and exploration of Adult Dataset

CS245, Data Science Foundation, Chaojun Lu, Autumn 2017

叶泽林 515030910468

1 Introduction

2 Approaches [1]

In sentiment analysis with lexicon-based polarity identification, the orientations of some opinion words might heavily depend on its context. For instance, consider the word *long*:

- The battery life is *long*.
- The time taken to focus is *long*.

Opinion word *long* for battery life is positive, while *long* for time to focus is negative. This problem is inevitable when using lexicon-based polarity identification, which ignores the context information.

3 Experiments

4 Conclusion

1. Apply the regular lexicon-based polarity identification method on the corpus and filter the adjectives with ambiguous meanings. The remaining ones are treated as absolute ground truth in later training.
2. Redefine a lexicon to be a word pair in the form of <*object, adjective*>, and apply the revised lexicon-based polarity identification method on it, where
   - The polarity of positive words: +1, negative word: -1, neutral words: 0.
   - The words classified by the regular lexicon-based polarity identification method in step 1, and the labeled lexicons in the seed dataset serve as ground truth when training.
   - When aggregating opinions, the equation

$$score(f_i, s) = \sum_{op_j \in s} \frac{op_j.so}{d(op_j, f_i)}, \tag{1}$$

     where op_j is an opinion word (either a ground truth or a induced lexicon) in s, and $d(op_j, f_i)$ is the distance between the closest words in op_j and f_i. $op_j.so$ is the orientation or the opinion score of op_j.
3. To make the model more accurate and robust, we further take the context and conjunctions(and, but, however ...) into consideration in long comments, which form links between sentences in each comment. Especially, the same word can appear in different sentences(e.g. "The battery life is long, but the time taken to focus is long"). Therefore, we can add the sentimental scores of nearby sentences with some weights to the target sentence factored by the conjunctive adverb. A Na$\acute{\text{i}}$ve design is that, if there are n conjunctive adverbs(e.g. *but*, *however*), op_j will be multiplied by a adjusting factor $(-1)^n$. A hand-engineered algorithm could possibly solve this problem, but it would be better to tackle the task by training a neural network.
4. After training, for each word we receive a score $\in [-1, 1]$, and some thresholds should be defined correspondingly to classify them into three classes.

5 Acknowledgement

References

[1] Sergey Levine, Peter Pastor, Alex Krizhevsky, and Deirdre Quillen. Learning hand-eye coordination for robotic grasping with large-scale data collection. In *International Symposium on Experimental Robotics*, pages 173–184. Springer, 2016.